MS # 6
MS pp 16
tab 0
FIG 0

CT **Introduction to Network Publishing**

rom
CA *Brewster Kahle*
Richard Koman
AA Wide Area Information Servers, Inc.

COFN Correspondence and requests for reprints should be sent to Brewster Kahle
Wide Area Information Servers, Inc.
1056 Noe Street
San Francisco, CA 94114
E-mail: Brewster@wais.com, Richard@ora.com

April 20, 1994

ABSTRACT

Network publishing is the integration of computer networks and traditional publishing that creates a basis for a new mechanism for organized information sharing. Computer networks can now support interactive text applications across many countries. Publishers have been exploiting computer technology to speed printed publications to market. Using computer networks for the distribution of work takes this trend to the next logical step. Based on the experience of the WAIS™, World Wide Web and Gopher systems on the Internet, this paper will propose the technical rationale for network publishing and suggest some of the components of a successful commercial system.

electronic publishing, networking, client/server computing

INTRODUCTION

Very few generations see a change in how people communicate with each other. When a new communications technology develops, all sorts of things change: industries form, groups of companies shift, methods of learning and sharing information change.

With the development of the printing press in the late 15th century, languages became regularized, people tapped the knowledge of the ancients and, more importantly, authors were able to spread their words far and wide. New types of writing flowered and literature was born. More recently, the telephone connected people in distant locations and allowed for the physical separation of offices and factories. While only businesses had access to the technology in the late 19th century, by the 1930s even rural homes were connected.

A similar ground-shifting technology is *network publishing*, the communication and distribution of information over wires. Network publishing has great potential to change -- and improve -- the flow of information. Network publishing opens doors for widespread access to all kinds of information, for new breeds of publishers, and for new business opportunities for traditional publishers.

Network publishing allows for inexpensive reproduction, targeted transmission and distributed control. These are the goals of the Wide Area Information Servers™ system, or WAIS. As the phrase suggests, network publishing comes out of the idea of a convergence of publishing and networks. The publishing industry is following a trend towards computerization, in which all elements of production are digital until the work is printed. Computer networks, notably the Internet, are now fast, inexpensive and highly distributed. This convergence of content and distribution lays the stage for network publishing. In the following sections, we will explore this convergence in more depth.

1.1 Publishing: Computerized Production

The publishing industry has undergone a series of technology shifts, most of which have come from the application of computer technology to such tasks at typesetting. The desktop publishing revolution -- based on Apple Macintosh computers, the Adobe PostScript page description language and graphics software -- accelerated the process. To understand the impact of this (and how it relates to network publishing), consider how the print publishing process has changed. In any

publishing operation, there are several pieces that must be integrated together to create a page that can be printed. The content must be written, the pages must be designed, the words must be typeset, black and white photographs must be converted to halftones, color photographs must be separated into sets of film. Then all the pieces are put together, film is shot of each page, the films are stripped together in a signature, a plate is made from the film, and finally the piece is printed.

The Macintosh and PostScript changed all that forever; entire industries were replaced in the process. With authors writing on computers and designers creating pages on screen, typesetting became inseparable from design. It was simply sucked into the machine. As the technology improved, more and more of these processes traditionally performed by skilled tradesmen were pulled into the computer as well. With current technology, all "prepress" functions can be done on desktop computers. But eventually the work is output to film for stripping, plate making and printing on paper.[1]

Computer Networks: Fast, Cheap, Distributed

Now that documents are created and stored on computers, a natural step is to distribute them in digital form. This requires a reliable digital transport that is inexpensive enough, fast enough and widespread enough to support text and graphics. The Internet has recently achieved these goals and is starting to be used as a distribution mechanism for network publishers. The cost of digital communications has dropped reflecting the drop in cost in terminal equipment and the rise in demand. Where text and graphics can now be shipped over computer networks, audio and video transmissions are not yet cost-effective. While phone lines are used to transmit fax pages, the slow speed means that users tend to receive these documents in the background for later reading. Internet speeds, on the other hand, are seven to 40 times faster. This is fast enough for users to browse, search, and scan text and business graphics. As the speeds go up, color graphics, audio, and video will be practical, but for now, network publishing is centered on text and simple graphics.

With an inexpensive system that is fast enough, the remaining question is: Does the network connect enough users? A recent study indicated that some 20 million people use Internet e-mail and over a million users are interactively connected,

[1] There are many positive things about printed communications but one problem is cost. Besides the cost of the substrate, paper must be warehoused, shipped, mailed, etc.

with usage doubling every nine months.[2] While the academic and research communities make up the current base of Internet users, the greatest growth is in the international and commercial sectors.[3] For publishers that want to address the markets connected to the Internet, that is enough coverage to support the first businesses.

Therefore the Internet is a viable distribution model for network publishing and it (or other similar networks) will likely prosper. The history of the Internet -- from Defense Department research project to education-and-research network to commercial backbone -- is the story of the successful migration of a scalable technology.

New Kinds of Publishers

Network publishing will turn many kinds of organizations into international publishers. Everyone who is in the business of providing information to an audience, whether for free or for some payment, may become a network publisher: government agencies, corporations, libraries, individual writers and artists, as well as publishers of magazines, newspapers and books. The first wave of network publishers -- and this is already happening -- is comprised of organizations who are looking for less-expensive ways of distributing free information. These include corporations, government agencies, university libraries and catalog publishers. Sun Microsystems is an example of this kind of network publisher. Sun uses the Internet to distribute technical marketing materials at a much lower cost and greater timeliness than could be done by printing and mailing these materials.

A second wave is comprised of newspaper, magazine, journal and book publishers who are "republishing" their content for networks. Dow Jones and Encyclopædia Britannica are two "traditional" publishers intent on making money by network publishing. Because traditional publishers already have in place a system for gathering, editing and presenting information, publishers can easily "repurpose" data collected for the primary business function. While many publishers are using online services like CompuServe and America Online to publish electronically, these companies are publishing directly on the Internet in order to maintain the profitability of their network publishing business. "The main reason we are doing it ourselves is that you just can't make any money licensing your content," Joseph J.

---

[2] Internet Society, CNRI, Reston, Virginia.

[3] ibid.

Esposito, president of Encyclopædia Britannica North America said in a *New York Times* article on the Britannica service.[4] "If you do believe that content is king, it's rather unfortunate that so many of the content providers have put themselves in a position where they're held hostage to the online services."

The third wave will occur when works are created directly for this interactive environment. People will take advantage of the fact that anyone can be a publisher: all that is needed is a computer, a telephone and something to say. Individuals will share ideas and work with others in a way not possible before. Network publishers will be able to find an audience and readers will be able to find compelling documents.

REQUIREMENTS OF NETWORK PUBLISHING

Information -- in the form of newspapers, magazines, television and telephone services -- is flooding into people's lives. Yet discussion of the information superhighway causes many people to say, "I can't deal with the information I get now. What do I want with more information?" This "information anxiety" signals that navigation and information tools are not good enough yet.

Readers need to be able to search and find the information they need without being exposed to information they don't need, to browse and explore other information when they have the time and inclination, and stay up-to-date by having new information delivered. Finally, they will want all of these information retrieval techniques integrated into a single interface. These are the goals of WAIS as a network publishing system: easy and efficient navigation, the development of a wide and varied community of users, and the ability to publish both free and for-pay information.

Searching

The primary tool in network publishing is the ability to ask for information on a topic and quickly get a response of relevant documents. The method should be intuitive, familiar, easy and fast. This is currently available with WAIS.

Browsing

Playing around in the information is crucial because it lets us understand the breadth of the information. Browsing, shopping, and exploring are required before we

---

[4] John Markoff, "Britannica's 44 Million Words Are Going On Line," The New York Times, February 8, 1994

even know what we can search for. In this way we will find new topics that we didn't know were interesting before. Browsing Internet resources is currently available through the Gopher and World Wide Web systems.

Updating

Staying up to date with our current interests can be difficult unless a steady stream of filtered information is automatically presented. If this is not done well enough, then we just won't find the time to read it. This process is similar to newspaper production, in which a staff of reporters and editors filter and interpret information, present it in an appealing way on a page, and the daily paper is delivered to subscribers' doorsteps. WAIS provides an infrastructure for agenting technology, which will deliver this ability to passively receive new information. See the agents section below for a further discussion of agents.

Information Integration

Users will require seamless access to all information -- personal, corporate and published -- ideally with the same tools. This breadth of information will probably not be in one library or database but rather include one's own files, enterprise databases, and many outside sources. Searching must be easy and intuitive even through the mountains to search through are large and unorganized.

These four processes address users' needs that are now primarily handled in print and telephone communications.

How does Network Publishing Deliver these Goals?

The goal is to combine the right network tools into a coherent whole that can be used over a wide area by millions of users. Moving through gigabytes of information efficiently requires advanced PC products, digital networks and a myriad of information sources to choose from. One system that is rapidly evolving towards that goal is WAIS. The pieces that make this possible are:

- Easy information navigation
- Client-server architecture
- Agent technology
- Security measures

3. INFORMATION SERVER AS A TOOL FOR FINDING INFORMATION

The primary requirement of a network publishing system is that it let users find the most relevant references out of a collection of 10,000 to one million documents. For this we need "information servers" -- smart servers that store large amounts of information -- from one to 100 gigabytes -- and support thousands of users. An information server needs as many clues as possible as to what the user likes and dislikes and should give as much feedback to the user about what it contains and what it can be used for.

Typically, users approach a database in three ways. First they may want to browse through the database, learning about its organization and scope. Then they will tend to search for information on several different topics. Finally, they will narrow to a very specific search. There are three principal ways that servers provide the searching capability: natural language searches, relevance feedback and Boolean searches.[5]

3.1 Natural Language

Writing a query should be no more difficult than asking a question. Natural language lets the user say "I'm looking for this" or "What do you know about that?" Natural language queries can include any number of extraneous words, can be in question or statement form, and require no special syntax, case sensitivity or mathematical symbols. This is the beginning of the dialog, a broad question answered by a number of possibly relevant documents.

3.2 Relevance Feedback

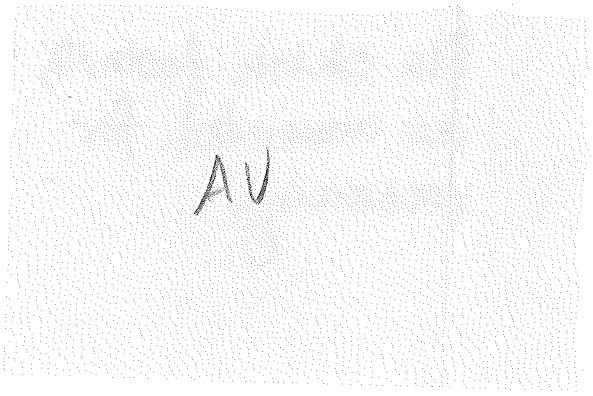
The conversation continues with relevance feedback when the user picks one of the returned documents, or a section of one document, and says: "I like that one -- find me more like that one."[6] This is a powerful mechanism for moving through large collections of information by finding documents that are "linked" to the current one. Rather than static hypertext links, however, these links can be dynamically created based on what the user has liked in the past.

---

[5] Some commercial Boolean systems are Dialog and Mead Data; Boolean and natural language systems are available from Fulcrum, Personal Librarian Software, Conquest; systems that include relevance feedback Thinking Machines Corporation verity, WAIS Inc.

[6] Gerard Salton and M. McGill. 1983. *Introduction to Modern Information Retrieval.* McGraw-Hill, New York.

Pls check footnote, as amended, for meaning.


AV

Boolean and Fielded Searches

When the user has narrowed in on the desired information, it is often fruitful to do very specific searches using Boolean and fielded search parameters. A certain amount of training and sophistication is required to make good use of these techniques but they can be quite powerful.

Future Searching Technologies

Future searching systems will handle multi-media, so that users can search on pictures as well as text, handle multiple languages, learn from past user feedback, etc.

CLIENT-SERVER TECHNOLOGY: THE POWER OF THE DESKTOP

Client-server frees users from the shackles of the mainframe. They can interact with servers using desktop computers, laptops, "personal digital assistants" and, maybe someday, home game machines. Client-server technology puts the control in the user's hands, by exploiting the power of the user's computer to provide more functionality and efficiency. The chief advantages of client-server for network publishing are graphical user interfaces (GUIs), integration with other applications and advanced display modes.

Graphical User Interfaces

Graphical user interfaces typically features icons and windows, thus hiding complexity and increasing ease-of-use and efficiency. In a client-server environment, the local computer controls the user experience and the server provides fixed services. By contrast, in a remote windowing system, the server controls the entire user experience. Examples of this are America Online and Mosaic.

Ease of use

Applications using icons and menus have been shown are 35% faster to use than a similar character-based program. GUI users were less fatigued and were found to explore and teach themselves the capabilities of the application.[7]

[7] Dekkers L. Davidson. 1990. *The Benefits of the Graphical User Interface.* Temple, Barker & Sloane.

Integration with the Desktop

Client-server allows users to bring external information into other local programs such as word processors, spreadsheets and image editors. It is also possible to add searching functions directly into these programs, so, for instance, a writer could search for a specific document, download it, display it and edit it, all without leaving the word-processing program. Or a designer could search the network for an appropriate stock photo and add it to the design, all within one page-layout program.

The Future of Client-Server: Agent Technology

In the future, client-server technology will be exploited to develop agents that will serve as alter egos. Agents will be able to ponder indications of a user's preferences and act accordingly. A user's computer will know what its owner reads and doesn't read, to whom messages are sent, whose messages are ignored, etc.[8]

Automating some of the information collection tasks can help find relevant information from thousands to tens of thousands of sources. Given the power of desktop machines and a protocol that allows for machine automation, we have the pieces needed to create these searching automatons. On the Internet, there are literally thousands of information servers, so a system of robots would be useful.

While the word "agents" suggests a human capability similar to a secretary or research assistant, the current technology is in its infancy. The precursors are present however: a growing body of quality information, computer-to-computer protocols that can support agents, multi-tasking operating systems on the desktop, digital networks and most importantly a discerning user population.

Today experimental agents are starting to perform the following tasks:

- Ask many servers a question on behalf of a user and track the user's actions in response to the answers.
- Scour the world (within a budget) to find new sources.
- Work 24 hours a day finding information.
- Format "personal newspapers" for users to read off-line on portable machines.

---

[8] Commercial systems include Apple's Rosebud project which became AppleSearch (Apple Computer, Cupertino California), and Relevant Personal Digital Newspaper by Ensemble Inc, Menlo Park California.

- Gossip with other clients to share information.

Organizations like Xerox's Palo Alto Research Center, Massachusetts Institute of Technology, General Magic and Ensemble are working to make these capabilities available in the next few years. By the time users start to use these automating processes, they may not even be aware of them.

NAVIGATING A SEA OF SERVERS: UBIQUITOUS PROTOCOLS

While client-server has significant value within an organizational LAN, it really shines in a wide-area network like the Internet. When a good protocol is in place, network users can access multiple servers in a single search, talk to many different kinds of servers in the same way, access personal, organizational and published (wide-area) information in an integrated fashion, use sophisticated clients, and use the network as a reference source with such tools as a directory of servers.[9] To provide all of this, the protocol must be flexible, extensible, standard and, of course, good enough.

Flexible

The protocol should operate on all computers -- from desktop personal computers to supercomputers; it should allow for searching of many data types -- not only text but also maps, DNA structures, and other unusual data types. The protocol should support any search syntax. Finally, it should allow clients to gossip with one another about their discoveries.

Extensible

An extensible protocol can grow and add new features without going through a long standardization process.

Standard

The protocol should be based on non-proprietary, international standards, so companies can compete but still be interoperable.

---

9 Protocol committees that are working on relevant network standards include: Internet Engineering Task Force, National Information Standards Organization, International Standards Organization, OSI Working Group on Library Applications. Document formats standards are set by other groups and companies.

Good enough

To be good enough, the protocol must be able to handle the current set of needs and be able to retrieve any kind of data, including text, graphics, sound and video.

SECURITY IN A WORLD WIDE NETWORK

Security systems in a network publishing system restrict unwanted access to documents based on the user's identity. Users ask the server for data and they are allowed or refused access according to the commands of the information provider. Unlike services that "broadcast" files (like NetNews or CD-ROM distribution), in a network publishing system, documents are only copied when a user requests them. Thus the publisher controls the distribution of the work.

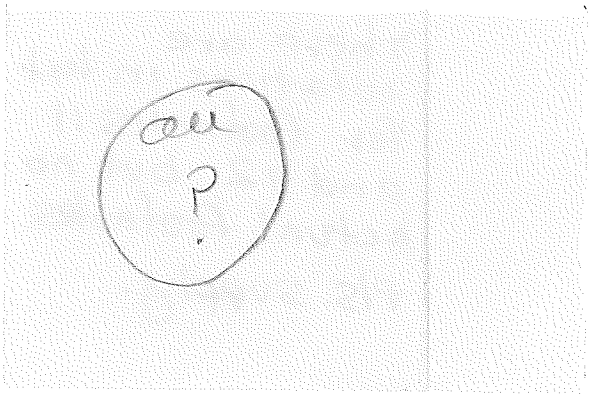
The primary security concerns in a network publishing environment are privacy, theft and viruses.

Privacy

Network publishing brings up new issues of privacy because most users are unaware that their actions can be recorded. For instance, WAIS, Gopher and World Wide Web all generate usage logs for the server. These logs tell the server administrators who searched for what information and what files were downloaded. While these logs provide valuable information that helps information providers improve their services, it is also possible that information could be sold to third parties. Encryption can protect information during transmission by encoding messages so they can only be read by the intended recipient. But this is only part of the answer. The network publishing community also needs to develop rules of conduct for handling this information.

Theft

In this context, theft refers primarily to improperly reselling information owned by someone else, or giving away copies of something that is being sold. Network publishing systems are attempting to make it easy for users to act legally by making "pointers" to the original data. Someone who wanted to include an article from a for-pay online magazine could simply construct a pointer that would bring the user to the site of the legitimate publisher of that document.


P

There are notations in text for footnotes 11 and 12 but no actual footnotes pls supply.

Another concern is that someone who had not paid for access to a for-pay server would be able to break in, thus depriving the provider of income. This possibility is effectively handled by authentication procedures.

Viruses and Security Breaches

In the WAIS system, users do not actually log in to the server; rather, they search through a read-only application layer protocol (Z39.50[10]). Thus there is no risk of information on the server being modified or of the server being contaminated by viruses.

Current Security Techniques

Securing information involves a balance of ease-of-use and protection. It would be prohibitively difficult for a user to remember different passwords for every server contacted; on the other hand, giving a user a single password for many information servers would invite abuses. Hardware encryption devices, like the proposed Clipper, are far more difficult to distribute than a software solution. Older security systems that allow two systems to communicate because they each know the same secret key[11] are difficult to extend to a system where there are thousands of servers and millions of users. Two new systems have been developed to address these problems -- public key and Kerberos -- and both are starting to be used in the WAIS environment.

Public key technology offers all the right pieces: privacy, scalability, authentication and digital signatures. The catch is it requires licensing from a private company. This has not stopped implementation but it has slowed dissemination. Kerberos, from MIT[12], does not require licensing, but requires a hierarchy of authorities to validate connections. All in all, the public key and Kerberos technologies offer strong security measures for network publishing, but the dissemination of the infrastructure will take awhile.

BILLING AND PAYMENT MODELS

Making it possible for small producers, as well as large, to be compensated for their work offers opportunity for continued growth in the Internet. Every company,

---

[10] Z39.50-1992 Information Retrieval Service and Protocol (ANSI/NISO).

every academic department, every family should be able to publish on this network. Some will want to be compensated. Facilitating this future industry is one of the goals of the WAIS system.

Collecting the customer usage information is technically not difficult, but many questions about how the business will develop remain unanswered. There are many possible billing structures: subscription, site-license subscriptions, pay-per-article, advertising-supported, and many others. While print publishing broke up into separate industries -- writing, publishing, printing, distributing, and retailing-- the network publishing business might evolve differently. In the current stage, the goal is to reach a critical mass of quality services that readers are willing to pay for.

Current businesses that are making money on the Internet include connectivity providers, hardware providers, telecommunications companies, book publishers and consultants. After the plumbing is done, then interactive information services such as magazines, games and performance events can proceed on the networks.

Some Internet information service providers (such as WAIS Inc., Bunyip and Pandora) are just starting to make money. These businesses typically take content from a publisher and retarget it for the networks. Some publishers (such as Encyclopædia Britannica) operate the systems themselves, but niche service bureaus offer expertise and economies of scale.

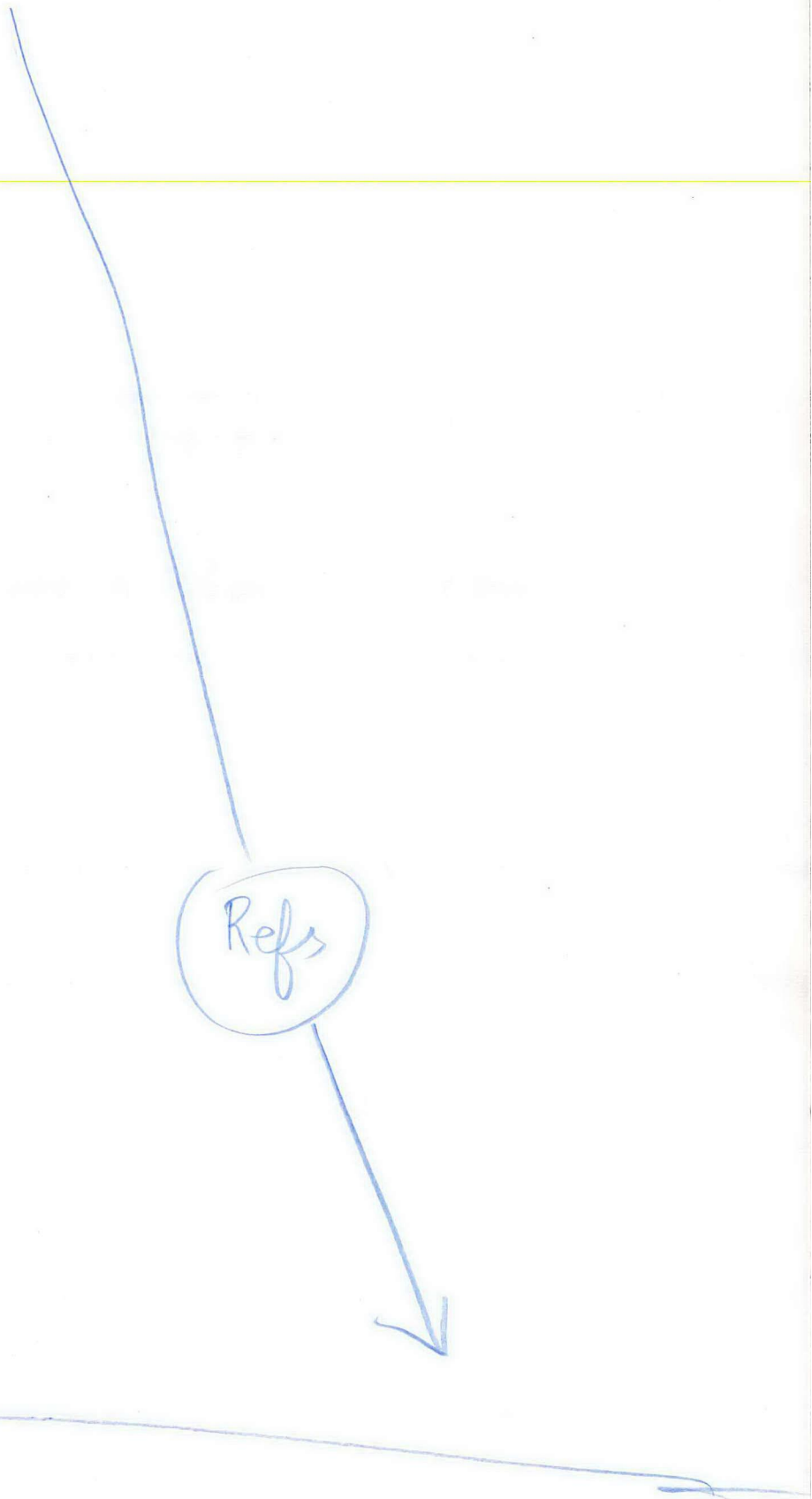
How the customer will pay for information access is another open question. End-users might be billed for single subscriptions, but more likely connectivity providers (such as regional Internet providers and online services) will act as middle-men to centralize billing.

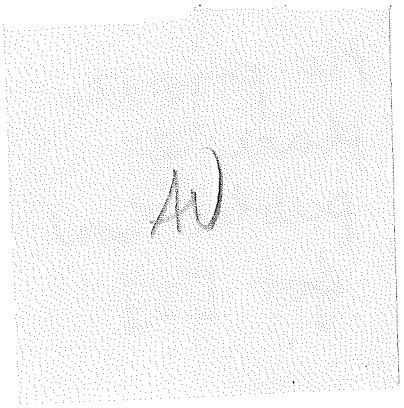
8. CONCLUSION

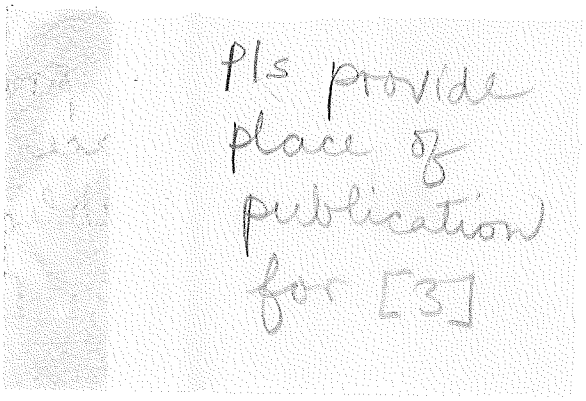
Weaving the network publishing elements together to make a usable system is the goal of WAIS. By incorporating a large number of servers and users, an open protocol for future growth and compensation mechanisms such a system can grow. So far the system has been useful on the Internet for search and retrieval, and WAIS resources have been blended into many other systems such as Gopher, World Wide Web, e-mail services and others. By January 1994, over 100,000 users have used WAIS and the number continue to grow.

Network publishing is not about saving trees or replacing books; it is about new relationships between publishers and readers, a fundamental shift in the way people

obtain information, and new forms of literature that will spring from a people unleashed to create and publish in an inexpensive new medium.






pls provide
place of
publication
for [3]

H1 RGF

References

all caps

6 (FN)

[1] J. Markoff, "Britannica's 44 million words are going on line", The New York Times, February 8, 1994.

8 (FN)

[2] G. Salton and M. McGill, Introduction to Modern Information Retrieval. McGraw-Hill, New York, 1983.

9 (FN)

[3] D. L. Davidson, The Benefits of the Graphical User Interface. Temple, Barker and Sloane, 1990.